A Novel Approach for Estimating the Omitted-Citation Rate of Bibliometric Databases With an Application to the Field of Bibliometrics

Fiorenzo Franceschini, Domenico Maisano, and Luca Mastrogiacomo
Politecnico di Torino, DIGEP (Department of Management and Production Engineering), Corso Duca degli Abruzzi, 24, 10129, Torino, Italy. E-mail: {fiorenzo.franceschini, domenico.maisano, luca.mastrogiacomo} @polito.it

One of the most significant inaccuracies of bibliometric databases is that of omitted citations, namely, missing electronic links between a paper of interest and some citing papers, which are (or should be) covered by the database. This paper proposes a novel approach for estimating a database's omitted-citation rate, based on the combined use of 2 or more bibliometric databases. A statistical model is also presented for (a) estimating the "true" number of citations received by individual papers or sets of papers, and (b) defining an appropriate confidence interval. The proposed approach could represent a first step towards the definition of a standard for evaluating the accuracy level of databases.

Introduction and Literature Review

The reliability of the information contained in any database is inevitably influenced by the presence of database errors (Kim, Choi, Hong, Kim, & Lee 2003). Bibliometric databases are no exception to the rule.

Historically, the dominant bibliometric database is Web of Science (WoS) by Thomson Reuters, which has been active for over 40 years and covers many of the high impact journals in each discipline (Garfield, 1979). In addition to this colossal multidisciplinary database, other database platforms dedicated to specific scientific fields have blossomed over the years: for example, MEDLINE and EMBASE in medicine, PsycINFO for psychology, MathSciNet in mathematics, etc. (Gavel & Iselid 2008; Michaleff et al., 2011). In 2004, the monopoly of WoS as a multidisciplinary database platform was broken by the advent of Scopus and Google Scholar (Vieira & Gomes 2009).

The problem of errors in bibliometric databases, particularly WoS, has occasionally been debated since the 1970s and 1980s. Just as an example, consider the contributions of Sweetland (1989) and Abt (1992). In the early years, reviewers, editors, and publishers did not always rigorously verify the correctness of the references listed by authors, probably because of the lack of practical tools for managing large volumes of bibliometric data. Also, article indexing was largely manual and therefore more prone to data-entry errors than at the present time. Over the past 10 years, the incidence of bibliometric database errors has been decreasing, probably due to the systematic employ of automatic tools for checking/correcting errors in cited article lists by editors and database administrators (Adam, 2002). Nevertheless, the problem is far from being solved, as proven by (a) several recent articles documenting the existence of different types of errors (see, for example, Jacsó, 2012), and (b) the fact that database providers constantly encourage users to report any noticed inaccuracy.

As observed (Jacsó, 2004, p. 40), the estimate of a database's error rate depends on the definition of the concept of error. According to a general definition, a measurement error is defined as "the difference between the value obtained by the measurement and the actual value of the measured quantity (the so-called 'nominal value' or 'true value', which is never known exactly)" (JCGM 200:2008, 2008, p. 22). This concept can be transferred to indicators based on the information contained in bibliometric databases; for example, it is possible to define the error relating to the total number of publications produced by a research institution in a given time window, the error related to the total number of citations received by a publication of interest, and so on. When estimating these errors, the most critical issue is probably knowing the true values. Ideally, it would be necessary to have an omni-covering and infallible database, that is to say a database with a complete coverage of the scientific

Received September 7, 2012; revised November 23, 2012; accepted November 24, 2012

© 2013 ASIS&T • Published online 21 August 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22898

literature and free from indexing errors. The information contained in this database would represent an "absolute reference" to determine the errors made by other imperfect databases. Unfortunately, such an ideal database does not exist and thus it is very difficult to estimate accurately database errors.

In the literature, most of the studies on database errors are based on the manual examination of sets of publications, so as to observe the consistency with the information/statistics contained in database(s). For example, the reader may consider the contributions of Buchanan (2006) and Li, Burnham, Lemley, and Britton (2010). In this case, the so-called true values are determined manually by analysts. However, these assessments—for obvious reasons—can never be exhaustive and error-free.

Another important issue when defining the concept of database error is that of database coverage. For example, at the moment WoS claims to cover more than 12,000 of the existing hundreds of thousands of scientific journals (Thomson Reuters, 2012a), and Scopus almost 19,000 (Scopus, Elsevier, 2012b). Google Scholar's coverage is probably higher (e.g., it also includes monographs, theses and dissertations, minor conference proceedings, etc.), but unfortunately the sources that are actually indexed by this database still remain a mystery (Bar-Ilan, 2008). Given the practical impossibility of having absolute coverage and given that the question of whether or not to cover some sources is a deliberate choice of bibliometric databases, our opinion is that the definition of error should be limited only to the publications included in the coverage statements of each database.

Another interesting point is the classification of possible error types. According to a classification by Buchanan (2006) there are two main categories: (a) author errors and (b) database mapping errors (for more details, see Table 1).

We emphasize that one of the main consequences of several error types is that of omitted citations, that is, citations that should be ascribed to a certain (cited) paper—being made by (citing) papers that are theoretically indexed by the database in use—but, for some reason, are lost. In other words, the link between citing and cited article is not established by the database.

TABLE 1. Classification of bibliometric database errors according to Buchanan (2006).

Error type	Author errors	Database mapping errors
Definition	Errors made by authors when creating the list of cited articles for their publication.	Failure to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to a data-entry error.
Examples	— Errors in name and initials of the first author, — Errors in publication title, — Errors in publication year, — Errors in volume number, — Errors in pagination.	— Transcription errors, — Target-source article record errors, — Cited article omitted from a cited-article list, — Reason unknown.

In addition to those in Table 1, there are other less frequent reasons why any of the cited articles are unlinked to the corresponding citing papers: for example, an internal page of the article had been cited, the database indexed a translation of the article (Nature, 2002; van Raan, 2005). According to the study by Buchanan (2006), which is based on a limited number of articles, omitted citations are likely to be around 5–10% of the total number of "true" citations.

The aim of this paper is to propose a novel approach to estimate the incidence of omitted citations in bibliometric databases. The method is based on the comparison of "overlapping" citation statistics concerning the same set of papers of interest, but provided by two or more different databases. In the absence of an absolute reference—that is, the true number of citations received by a paper of interest—this redundancy of information allows a reasonable estimate of the degree of inaccuracy of a database with respect to others. The procedure will be tested using a set of papers indexed by Scopus and WoS, as they both are multidisciplinary databases with well-known lists of covered sources.

Although there are numerous studies comparing different bibliometric databases—for example, the contributions of Bakkalbasi, Bauer, Glover, and Wang (2006), Gavel and Iselid (2008), Archambault, Campbell, Gingras, and Larivière, (2009)—they almost entirely focus on investigating their coverage. There are also several studies on the inaccuracy of various databases, from which it emerges that Google Scholar is significantly dirtier than other databases—for example, see the contributions of Meho and Yang (2007), Jacsó (2009), and Franceschini and Maisano (2011). Unfortunately, such information is rarely accompanied by quantitative data. The main strengths of this novel approach are that (a) it can give a quantitative estimate of the omitted-citation rate of a database or a portion of it and (b) it can be automated.

The remainder of this paper is structured in four sections. Presentation of the Method describes the proposed method in detail, highlighting its assumptions. In Application Example, the method is implemented based on a limited sample of papers and considering the WoS and Scopus databases. The Statistical Model section presents a statistical model that—given the total citations counted by a database for a paper or set of papers—allows the estimation of the true number of citations (taking into account the omitted-citation rate) with an appropriate confidence interval. We conclude by highlighting the main implications, limitations, and original contributions of this article.

Presentation of the Method

Introductory Example

Before providing a comprehensive description of the suggested method, we present an introductory example to illustrate how it works.

TABLE 2. Citation statistics relating to an article of interest (the paper with ID No. J1.33 from table in supporting information), according to Scopus and WoS. The nine citing articles highlighted in gray belong to sources purportedly covered by both databases.

	Citations in	
Citing article no.	Scopus	WoS
1	✓	
2		✓
3	✓	✓
4	✓	✓
5	✓	
6	✓	
7	✓	
8	✓	
9	✓	
10	✓	
11	✓	
12	✓	✓
13	Omitted	✓
14		✓
15	✓	✓
16	✓	✓
17	✓	
18	Omitted	✓
19	✓	✓
20		✓
21	✓	Omitted
22	✓	
23	✓	
24	✓	
25		✓
Total	19	12

Let us consider a paper of interest—for example, the article with ID No. J1.33 from the supporting information—indexed by Scopus and WoS.[1] On 27th July 2012, the number of citations received by this paper are 19 in Scopus and 12 in WoS. As the level of coverage by Scopus is generally higher than that of WoS, this disparity is not surprising (Bar-Ilan, 2008).

The union of the citations recorded by the two databases (see the first column of Table 2) is a total of 25 citations. Among the citing articles, only nine belong to sources (i.e., journals or conference proceedings) officially covered by both databases[2] (highlighted in gray in Table 2). Focusing on the nine overlapping citing articles, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/9 \approx 22.2\%$ in Scopus and $1/9 \approx 11.1\%$ in WoS.

The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

Detailed Description

The procedure of automatic analysis of omitted citations, which is based on the use of two or more bibliometric databases, is described in the following steps (see the scheme in Fig. 1):

1. Identify a sample of scientific publications, for example, papers produced by groups of researchers, research institutions or nations. Among these papers, identify those indexed by at least two databases; this can be done by querying the databases in use with the DOI code and the full title of each paper. Documents not indexed by at least two of the databases in use, that is, those outside the so-called database "intersections" or "overlaps" are excluded from the sample.
2. Consider each (*i*-th) paper of the sample individually.
    2.1. Identify the total citing papers relating to each *i*-th paper of the sample, that is, the union of citing papers detected by all databases.
    2.2. Consider the perspective of a single (*j*-th) database of those in use.
    2.2.1. Identify the "theoretically overlapping" citing papers relating to the *i*-th paper of the sample. They are defined as the portion of the citing papers identified at point (2.1) that are issued by journals officially indexed by (a) the *j*-th database and (b) at least one other of the databases in use. For example, in Table 2, the "theoretically overlapping" citing papers relating to Scopus are those highlighted in gray, since they are issued by journals covered by Scopus and one other database (i.e., WoS). This verification can be done by using the article issue year and the ISSN code of the journal, seeing whether articles are included in the corresponding annual official list of documents covered by the database. For example, as regards WoS, the *Journal Citation Reports* can be used (Thomson Reuters, 2012b), while, as regards Scopus, the official list available on the website can be used (Scopus Elsevier, 2012b).
    2.2.2. Among the theoretically overlapping citations (at point [2.2.1]), count the citations omitted by each database. A theoretically overlapping citation that does not occur in the database of interest is classified as omitted. For example, among the nine theoretically overlapping citing papers in the example of Table 2, only six are really overlapping, since two papers (i.e., the 13th and 18th) are omitted by Scopus, while one (i.e., the 21st) is omitted by WoS.

    (Repeat step 2.2. and substeps for all the databases in use)

    (Repeat step 2 and substeps for all the papers of the sample)

[1] The WoS database configuration included the following resources: *Citation Index Expanded* (SCI-EXPANDED) from 1970 to present, *Social Sciences Citation Index* (SSCI) from 1970 to present, *Arts & Humanities Citation Index* (A&HCI) from 1975 to present, *Conference Proceedings Citation Index—Science* (CPCI-S) from 1990 to present, *Conference Proceedings Citation Index - Social Science & Humanities* (CPCI-SSH) from 1990 to present.

[2] As a curiosity, the majority of the papers indexed exclusively by Scopus come from the INISTA (International Symposium on Innovation in Intelligent Systems and Applications) conference proceedings, which are not (yet) indexed by WoS.

DOI: 10.1002/asi

3. For each database, estimate the omitted-citation rate (p) relating to the sample of publications examined as:

$$p = \Omega / \Gamma, \tag{1}$$

Ω being the total omitted citations and Γ the total theoretically overlapping citations concerning the set of the papers of the sample.

Since it is founded on multiple queries over several databases, the proposed procedure is resource consuming. Fortunately, it can be automated.

To fully understand the potential/limitations of the procedure, it is convenient to focus on its simplifying assumptions:

- It is assumed that the omitted citations of different databases are statistically independent. Actually, to identify a citing paper omitted by a certain database, it is necessary that the same citing paper occurs in at least one of the other databases in use. Of course, the concurrent omission of a citing paper by all databases will prevent its detection, leading to an underestimation of p. However, based on the analysis of a relatively limited sample of (citing) papers, Buchanan (2006) shows that this hypothesis of independence is quite reasonable.
- It is assumed that the incidence of "phantom citations"—that is, erratic citations from papers that did not actually cite the target paper (Jacsó, 2008)—is negligible. According to our procedure, a phantom citation of a certain database—if it is (mistakenly) assigned to a paper that is supposed to be covered by other databases—may lead to an incorrect notification of omitted citation for these other databases. However, we note that the incidence of phantom citations is very low for most databases (Garcia-Pérez, 2010).
- As regards the sample of publications examined, p is estimated by using only a portion of the information available from the database of interest—that is, that on citing papers purportedly covered by at least one other database (see Figure 1). The results can be extended to the rest of the citations, upon the reasonable assumption that the incidence of omissions is similar.
- The extension of the value of p—estimated on the basis of a sample of papers examined—to the totality of the papers covered by the database of interest is a very delicate question. For instance, in the case p were presumably influenced by (a) scientific field and (b) age of the papers in the sample, the extension of its estimate to the whole database would be

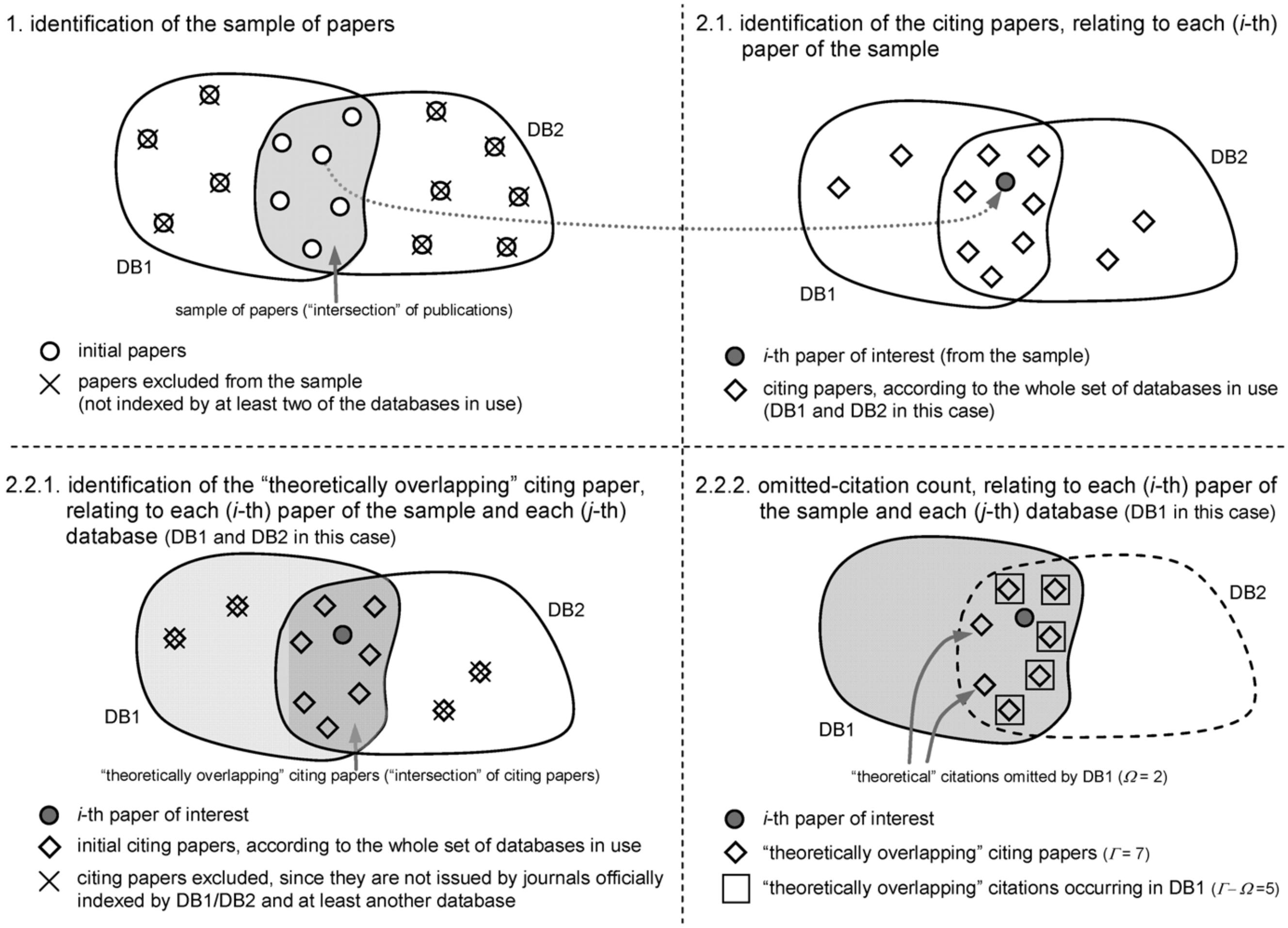
1. identification of the sample of papers
DB1
DB2
sample of papers ("intersection" of publications)
○ initial papers
× papers excluded from the sample (not indexed by at least two of the databases in use)
2.1. identification of the citing papers, relating to each (i-th) paper of the sample
DB1
DB2
● i-th paper of interest (from the sample)
◇ citing papers, according to the whole set of databases in use (DB1 and DB2 in this case)
2.2.1. identification of the "theoretically overlapping" citing paper, relating to each (i-th) paper of the sample and each (j-th) database (DB1 and DB2 in this case)
DB1
DB2
"theoretically overlapping" citing papers ("intersection" of citing papers)
● i-th paper of interest
◇ initial citing papers, according to the whole set of databases in use
× citing papers excluded, since they are not issued by journals officially indexed by DB1/DB2 and at least another database
2.2.2. omitted-citation count, relating to each (i-th) paper of the sample and each (j-th) database (DB1 in this case)
DB1
DB2
"theoretical" citations omitted by DB1 (Ω = 2)
● i-th paper of interest
◇ "theoretically overlapping" citing papers (Γ = 7)
□ "theoretically overlapping" citations occurring in DB1 (Γ – Ω =5)

FIG. 1. Main steps of the proposed method. Only two databases (i.e., DB1 and DB2) are considered for the sake of simplicity.

reckless. In order to shed more light on this point, future research will focus on a structured study of (non-)uniformity in the distribution of omitted citations within databases.

- The procedure can be readily applied to journal articles, but not as easily to other publication types—for example, book chapters, conference proceedings, monographs, etc.—for two basic reasons: (a) some of these publication types are not covered by databases, (b) lack of exhaustive official lists concerning the coverage of these document types. Incidentally, we note that Google Scholar cannot be used in our procedure because—apart from its large level of dirtiness (Jacsó, 2012)—it does not (yet) provide any official list of the documents covered.

Application Example

Feasibility of the proposed method was tested based on a sample of 343 papers issued by three scientific journals—that is, *Journal of the American Society for Information Science and Technology, Scientometrics* and *Journal of Informetrics*—in the year 2008. The analysis was carried out on 27th July 2012, using WoS and Scopus, and it was automated by an ad hoc software application. Total querying time was around 120 minutes.

Journals were identified by matching the journal title, ISSN code, and issue year with the information contained in the official lists of the journals covered by the databases (Thomson Reuters, 2012a; Scopus, 2012b). Papers in the sample and related citing articles were disambiguated by using the corresponding full titles and DOI codes.

Overall results of the analysis are synthesized in Table 3, while detailed results—that is, those at the level of individual papers of the sample—are reported in the supporting information. Out of 2,285 total theoretically overlapping citations at the intersection of the two databases, the number of omitted citations was 73 for Scopus and 128 for WoS; the corresponding values of p are $73/2{,}285 \approx 3.2\%$ and $128/2{,}285 \approx 5.6\%$, respectively. Although the sample of papers is relatively small, this result is in line with the estimates by Moed (2002) and Buchanan (2006). It is worth noting the lack of correlation ($R^2 \approx 0$) between the number of omitted citations at the level of individual papers (see supporting information). This is a confirmation of their statistical independence.

TABLE 3. Synthetic results relating to the application of the proposed method.

			Scopus		WoS	
Journal	N	Γ	Ω	p	Ω	p
(J1) *Journal of the American Society for Information Science and Technology*	188	1140	45	3.95%	61	5.35%
(J2) *Scientometrics*	125	934	22	2.36%	61	6.53%
(J3) *Journal of Informetrics*	30	211	6	2.84%	6	2.84%
Overall statistics	343	2285	73	3.19%	128	5.60%

Note. The sample is composed by the papers issued in 2008 by three scientific journals (in the first column). The following indicators are reported, both at journal and overall level: total number of papers (N), total number of theoretically overlapping citing papers (Γ), number of omitted citations (Ω), and omitted-citation rate ($p = \Omega/\Gamma$) of each database.

We note that these estimates of p should not be immediately extended to the entire databases or relatively large portions of them (e.g., research fields). Before doing this, it is essential to repeat the analysis using a more representative sample of papers and studying the influence of other factors—such as scientific field or age of the papers examined—on p, as we plan to do in the future.

A (manual) survey of the cited/citing papers showed that the most frequent reasons for omitted citations are database mapping errors due to target-source article record errors. Another rather frequent inaccuracy involves the article DOI string: For some articles it may happen that a character in the string is mistaken or missed by one of the databases. However, cross-references between the article full titles and DOI strings provided by multiple databases allow the identification of such inaccuracies, preventing them from generating erroneous article duplications, which may distort the analysis.

Statistical Model

For a generic i-th paper, the relationship between (a) the true citations (c_i^*, i.e., citations given by papers that are purportedly indexed by the database in use), (b) the real citations returned by the database (c_i), and (c) the citations omitted (o_i) by the database in use is modeled by:

$$c_i^* = c_i + o_i. \tag{2}$$

In this model, c_i is a known constant parameter related to the i-th paper. On the other hand, o_i will be estimated on the basis of the database omitted-citation rate (p) and treated as a random variable. Since c_i^* is an unknown parameter, it is replaced by its estimate $\hat{c}_i^*$. Being $\hat{c}_i^*$ a function of o_i, it will be treated as a random variable too.

The expected value and variance of $\hat{c}_i^*$ are respectively:

$$\mathrm{E}(\hat{c}_i^*) = c_i + \mathrm{E}(o_i), \tag{3}$$

$$\mathrm{V}(\hat{c}_i^*) = 0 + \mathrm{V}(o_i) = \mathrm{V}(o_i). \tag{4}$$

Let us now focus on the estimation of $\mathrm{E}(o_i)$ and $\mathrm{V}(o_i)$. The variable o_i can be modeled by a binomial distribution. Given (a) a generic i-th paper with c_i^* true citations and (b) the omitted-citation rate (p) related to articles homologous to the one of interest, the database's probability of omitting o_i citations is:

$$P(o_i) = \binom{c_i^*}{o_i} p^{o_i} (1-p)^{c_i^* - o_i}. \tag{5}$$

Since c_i^* is unknown, it can be replaced by $\mathrm{E}(\hat{c}_i^*)$, that is, its best estimate:

$$P(o_i) = \binom{\mathrm{E}(\hat{c}_i^*)}{o_i} p^{o_i} (1-p)^{\mathrm{E}(\hat{c}_i^*)-o_i}. \tag{6}$$

The expected value and the variance of the (random) variable o_i are respectively:

$$\mathrm{E}(o_i) = \mathrm{E}(\hat{c}_i^*) \cdot p, \tag{7}$$

$$V(o_i) = \mathrm{E}(\hat{c}_i^*) \cdot p \cdot (1-p). \tag{8}$$

Combining Equation (3) with Equation (7), it follows that:

$$\mathrm{E}(\hat{c}_i^*) = c_i + \mathrm{E}(\hat{c}_i^*) \cdot p. \tag{9}$$

From which it is obtained that:

$$\mathrm{E}(\hat{c}_i^*) = \frac{c_i}{1-p}. \tag{10}$$

Combining Equation (4) with Equations (8) and (10), it is obtained that:

$$\mathrm{V}(\hat{c}_i^*) = \mathrm{E}(\hat{c}_i^*) \cdot p \cdot (1-p) = c_i \cdot p. \tag{11}$$

Now we are able to describe $\mathrm{E}(o_i)$ and $\mathrm{V}(o_i)$ as functions of c_i.

Combining Equation 7 and Equation 10, it follows that:

$$\mathrm{E}(o_i) = c_i \cdot \frac{p}{1-p}. \tag{12}$$

As expected, o_i tends to increase linearly with c_i and—less trivially—it tends to increase more than linearly with p.

Combining Equation (4) and Equation (11), it follows that:

$$V(o_i) = \mathrm{V}(\hat{c}_i^*) = c_i \cdot p. \tag{13}$$

The variance of o_i grows linearly with c_i and p.

Let us now leave the perspective of a single i-th paper from the sample, in order to focus on multiple papers. Considering a set of P papers for which the database returns $C = \sum_{i=1}^{P} c_i$ total citations, the total number of omitted citations is:

$$O = \sum_{i=1}^{P} o_i. \tag{14}$$

The expected value and the variance of O can be obtained as follows:

$$\mathrm{E}(O) = \sum_{i=1}^{P} \mathrm{E}(o_i) = \sum_{i=1}^{P} c_i \cdot \frac{p}{1-p} = C \cdot \frac{p}{1-p} \tag{15}$$

$$\mathrm{V}(O) = \sum_{i=1}^{P} \mathrm{V}(o_i) = \sum_{i=1}^{P} c_i \cdot p = C \cdot p. \tag{16}$$

It was assumed statistical independence among the o_i values related to different papers of the sample.

Please note that Equations (15) and (16) could be obtained by applying the binomial probability distribution function to a group of (P) articles with $C^* = \sum_{i=1}^{P} c_i^*$ (unknown) total true citations. Precisely, for a group of papers with C^* total true citations, the database's probability of omitting O citations is:

$$P(O) = \binom{\mathrm{E}(\hat{C}^*)}{O} p^{O} (1-p)^{\mathrm{E}(\hat{C}^*)-O}, \tag{17}$$

being $\mathrm{E}(\hat{C}^*)$ the best estimate of the unknown parameter C^*.

Then, following this same logic, Equations (7) and (8) turn into Equations (15) and (16). Similarly, Equations (10) and (11) turn into:

$$\mathrm{E}(\hat{C}^*) = \frac{C}{1-p}. \tag{18}$$

$$\mathrm{V}(\hat{C}^*) = C \cdot p \tag{19}$$

Let us just focus on the practicality of the model developed in this section. Equations (10) and (11)—in the case of individual papers—and Equations (18) and (19)—in the case of multiple papers—may be useful to correct the citation statistics actually returned from a database (i.e., c_i or C values), compensating for omitted citations. As an example, suppose we query the Scopus and WoS databases to the total number of citations received by a group of three papers from the *Journal of the American Society for Information Science and Technology* (see Table 4a).

With regard to the ($C = 174$) total citations returned by Scopus, the mean and the variance associated with the estimate of C^* are obtained using Equations (18) and (19):

$$\begin{aligned} \mathrm{E}(\hat{C}^*) &= \frac{C}{1-p} = \frac{174}{1-3.19\%} \approx 179.7 \\ \mathrm{V}(\hat{C}^*) &= C \cdot p = 174 \cdot 3.19\% \approx 5.55 \end{aligned}. \tag{20}$$

For simplicity, the supposed-to-be-known p is the one relating to the sample of articles analysed before (in Table 3).

Since O follows a binomial distribution and $C^* = C + O$, then the lower and upper limits ($\hat{C}_l^*$ and $\hat{C}_u^*$, respectively) of a 95% confidence interval around $\mathrm{E}(\hat{C}^*)$ can be calculated by solving the following system of uncoupled equations:

$$\begin{cases} B[(\hat{C}_l^* - C) | \langle \mathrm{E}(\hat{C}^*) \rangle, p] = \dfrac{\alpha}{2} \\ B[(\hat{C}_u^* - C) | \langle \mathrm{E}(\hat{C}^*) \rangle, p] = 1 - \dfrac{\alpha}{2} \end{cases}, \tag{21}$$

TABLE 4. (a) Citations returned by Scopus and WoS relating to three papers issued in *Journal of the American Society for Information Science and Technology*; Article ID numbers refer to the papers reported in the list in supporting information table. (b) Results of the statistical model: $E(\hat{C}^*)$ is the best estimate of C^* and *CI* is the relevant 95% confidence interval.

(a) Citation statistics

	c_i	
Article ID No.	Scopus	WoS
J1.28	61	58
J1.79	97	93
J1.100	16	13
Total (*C*)	174	164

(b) Results of the Statistical Model

	Scopus	WoS
$E(\hat{C}^*)$	180	173
CI (binomial distribution)	[176, 184]	[168, 179]
CI (normal approximation to binomial)	[175, 184]	[168, 180]

being:

- *B[x|n,p]* the cumulative distribution function value at *x*, for a binomially distributed variable with parameters *n* and *p*;
- $\langle E(\hat{C}^*)\rangle = 180$, where $\langle\ \rangle$ denotes the rounding to the nearest integer;
- $\alpha = 5\%$, that is, the type-I error.

Regarding the data in Equation (20), the resulting confidence interval is $[\hat{C}_l^*, \hat{C}_u^*] \approx [176, 184]$. As a confirmation:

$$\begin{cases} B[2 \mid 180, 3.19\%] \cong 2.39\% \\ B[10 \mid 180, 3.19\%] \cong 97.5\% \end{cases}. \tag{22}$$

Similarly, it can be constructed a confidence interval regarding the citation statistics from WoS (see Table 4).

Since (1) $\hat{C}^*$ follows a binomial distribution (translated to the right by *C*) and (2) $E(\hat{C}^*)$ is usually large enough so that $E(\hat{C}^*) \cdot p \geq 5$ (Montgomery, 2005), it can be convenient to approximate the distribution of $\hat{C}^*$ by a normal distribution with mean and variance calculated in Equation (20). Therefore, a simplified way for determining a 95% confidence interval of $\hat{C}^*$ is:

$$E\left(\hat{C}^*\right) \pm 2 \cdot \sqrt{V\left(\hat{C}^*\right)} = \frac{C}{1-p} \pm 2\sqrt{C \cdot p} = 179.7 \pm 4.7. \tag{23}$$

Rounding these limits to the nearest integers, the result that is obtained (i.e., [175, 184]) is in line with that obtained when using the binomial distribution. The same applies to the citation statistics from WoS (see Table 4b).

Obviously, the proposed correction model can be applied to groups of publications or even individual publications (using Equations [10] and [11] instead of Equations [18] and [19]). In the our opinion, this model is the starting point for more realistic estimates of any bibliometric indicator based on citation statistics (e.g., the h-index, the average citations per paper, etc.).

Conclusions

This paper presents a novel procedure for estimating the omitted-citation rate of (portions of) bibliometric databases, based on the combined use of two or more databases. The basic logic is that the mismatch between the citations occurring in one database and another is evidence of possible errors/omissions.

The greatest strength of the procedure is that it can be automated and, hence, applied to a large number of papers so as to provide relatively robust, transparent, and repeatable estimates. This procedure may be of interest to database users and professionals, as it allows a comparison between competing databases. Also, it could become a standard for evaluating the accuracy level of databases and detecting/correcting their omitted citations automatically. Of particular interest is the statistical model for correcting the citation statistics given by a database.

Some limitations of the procedure may include: (a) for each database, it is necessary to have an official list of documents supposedly covered; (b) although it is automated, the suggested procedure is somewhat time-consuming; and (c) it can be influenced by nonuniformity in the distribution of citations omitted by database documents. We should note that while the fact that omitted citations are distributed uniformly among database documents might seem reasonable for a relatively homogeneous set papers (e.g., same scientific field or age), it could be at least doubtful in the general case.

Future developments of this research concern the implementation of the proposed procedure on a larger sample of articles, so as to provide a more robust estimate of *p* and investigate the possible influence on it of two factors: (a) scientific field, and (b) age of the publications examined. Should systematic differences in the value of *p* be detected, the proposed model for estimating the true citations would remain valid, provided that the unique global *p* of a database is replaced with local estimates of it, that is, using samples of articles homologous to that/those of interest.

Furthermore, the statistical model may be integrated with informetric models of indicators based on citation statistics (e.g., the h-index), in order to analyze the effect of omitted citations on them (Franceschini and Maisano, 2010; Franceschini et al., 2012a–c).

References

Abt, H.A. (1992). What fraction of literature references are incorrect? Publications of the Astronomical Society of the Pacific, 104(Mar 1992), 235–236.

Adam, D. (2002). Citation analysis: The counting house. Nature, 415(6873), 726–729.

Archambault, E., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing of science bibliometric statistics obtained from the Web of

Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320–1326.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Goole Scholar, Scopus and Web of Science. Biomedical Digital Libraries, 3,7. Retrieved from: http://www.bio-diglib.com/content/3/1/7 [July 2012].

Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. Scientometrics, 74(2), 257–271.

Buchanan, R.A. (2006). Accuracy of cited references: The role of citation databases. College & Research Libraries, 67(4), 292–303.

Franceschini, F., & Maisano D. (2010). Analysis of the Hirsch index's operational properties. European Journal of Operational Research, 203(2), 494–504.

Franceschini, F., & Maisano, D. (2011). Influence of database mistakes on journal citation analysis: Remarks on the paper by Franceschini and Maisano, QREI (2010). Quality and Reliability Engineering International, 27(7), 969–976.

Franceschini, F., Galetto, M., Maisano, D., & Mastrogiacomo, L. (2012a). An informetric model for the *success*-index. To appear on Journal of Informetrics.

Franceschini, F., Galetto M., Maisano D., & Mastrogiacomo L. (2012b). The success-index: An alternative approach to the h-index for evaluating an individual's research output. Scientometrics, 92(3), 621–641.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2012c) The effect of database dirty data on h-index calculation. To appear in Scientometrics, DOI 10.1007/s11192-012-0871-x.

Garcia-Pérez, M.A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h-indices in psychology. Journal of the American Society for Information Science and Technology, 61(10), 2070–2085.

Garfield, E. (1979). Citation indexing. Its theory and application in science, technology and humanities. New York: John Wiley & Sons.

Gavel Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. Online Information Review, 32(1), 8–21.

Jacsó, P. (2004). The future of citation indexing: An interview with Eugene Garfield. Online, 28(Jan./Feb. 2004), 38–40.

Jacsó, P. (2008). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. Library Trends, 56(4), 784–815.

Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. Online Information Review, 33, 376–385.

Jacsó, P. (2012). Grim tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: Reflections on Vanclay's criticism. Scientometrics, 92(2), 325–354.

JCGM 200:2008. (2008). VIM—International vocabulary of metrology—Basic and general concepts and associated terms (VIM). International Organization for Standardization, Geneva, Switzerland.

Kim, W., Choi, B.J, Hong, E.K., Kim, S.K., & Lee, D. (2003). A taxonomy of dirty data. Data Mining and Knowledge Discovery, 7(1), 81–99.

Li, J., Burnham, J.F., Lemley, T., & Britton, R.M. (2010). Citation analysis: Comparison of Web of Science, Scopus, Scifinder, and Google Scholar. Journal of Electronic Resources in Medical Libraries 7(3), 196–217.

Meho, L.I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology 58(13), 2105–2125.

Michaleff, Z.A., Costa, L.O., Moseley, A.M., Maher, C.G., Elkins, M.R., Herbert, R.D., . . . (2011). CENTRAL, PEDro, PubMed, and EMBASE are the most comprehensive databases indexing randomized controlled trials of physical therapy interventions. Physical Therapy, 91(2), 190–197.

Moed, H. (2002). The impact-factors debate: The ISI's uses and limits. Nature, 415(6873), 731–732.

Montgomery, D.C. (2005). Introduction to statistical quality control, 5th ed. Hoboken, NJ: Wiley.

Nature. (2002). Editorial: Errors in citation statistics. Nature, 415(6868), 101.

Scopus Elsevier. (2012a). FAQs. Available at: http://www.info.sciverse.com/scopus/scopus-training/faqs

Scopus Elsevier. (2012b). Scopus content coverage. Available at: http://www.scopus.com

Sweetland, J.H. (1989). Errors in bibliographic citations: a continuing problem. Library Quarterly, 59(4), 291–304.

Thomson Reuters. (2012a). ISI Web of Knowledge. Available at: http://thomsonreuters.com/

Thomson Reuters. (2012b). Journal Citation Reports. Available at: http://admin-apps.webofknowledge.com/JCR/JCR

Van Raan. (2005). For your citations only? Hot topics in bibliometric analysis. Measurement, 3(1), 50–52.

Vieira, E.S., & Gomes, J.A.N.F. (2009). A comparison of Scopus and Web of Science for a typical university. Scientometrics, 81(2), 587–600.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

TABLE S.1. Results relating to the application of the proposed procedure at article level. The sample is composed by the articles issued in 2008 by three scientific journals – that is, *Journal of the American Society for Information Science and Technology* (J1), *Scientometrics* (J2) *and Journal of Informetrics* (J3). For each article of the sample, the following data are reported: article ID No., number of citations (c_i) returned by each database, number of theoretically overlapping citations (γ_i), and number of citations omitted (ω_i) by each database. Articles are sorted in ascending order according to their DOI code (shown in the digital supplementary file).